WPS5207

Policy Research Working Paper 5207

A New Approach to Producing Geographic Profiles of HIV Prevalence

An Application to Malawi

Oleksiy Ivaschenko
Peter Lanjouw

The World Bank
Europe and Central Asia Region
Human Development Economics Unit
&
Development Research Group
Poverty and Inequality Team
February 2010

Policy Research Working Paper 5207

Abstract

Sub-national estimates of HIV prevalence can inform the design of policy responses to the HIV epidemic. Such responses also benefit from a better understanding of the correlates of HIV status, including the association between HIV and geographical characteristics of localities. In recent years, several countries in Africa have implemented household surveys (such as Demographic and Health Surveys) that include HIV testing of the adult population, providing estimates of HIV prevalence rates at the sub-national level. These surveys are known to suffer from non-response bias, but are nonetheless thought to represent a marked improvement over alternatives such as sentinel surveys. At present, however, most countries are not in a position to regularly field such household surveys. This paper proposes a new approach to the estimation of HIV prevalence for relatively small geographic areas in settings where national population-based surveys of prevalence are not available. The proposed approach aims to overcome some of the difficulties with prevailing methods of deriving HIV prevalence estimates (at both national and sub-national levels) directly from sentinel surveys. The paper also outlines some of the limitations of the proposed approach.

This paper—a joint product of the Human Development Economics, Europe and Central Asia Region; and Poverty and Inequality Team, Development Research Group—is part of a larger effort to publish policy-relevant research. Policy Research Working Papers are also posted on the Web at http://econ.worldbank.org. The authors may be contacted at oivaschenko@worldbank.org and planjouw@worldbank.org.

A New Approach to Producing Geographic Profiles of HIV Prevalence: An Application to Malawi

Oleksiy Ivaschenko and Peter Lanjouw[1]

Key words: HIV prevalence, census, GIS, sentinel data, small area estimation, bootstrap.

[1] The affiliation and contact details of the authors are as follows: Oleksiy Ivaschenko, Economist, Europe and Central Asia Human Development Department, The World Bank, email: oivaschenko@worldbank.org; Peter Lanjouw, Research Manager, Research Manager, Development Research Group, The World Bank, email: planjouw@worldbank.org. The authors would like to thank Livia Montana for providing generous support for this research. We are also very grateful to all those who provided comments and suggestions on the earlier version of this paper. All the views expressed in this paper are solely those of the authors, and should not be attributed to the World Bank group.

1. Introduction

Collecting data on HIV/AIDS is notoriously difficult. As a result, common data sources in most countries do not yield representative sub-national prevalence estimates. As in most Sub-Saharan African countries, Malawi monitors HIV prevalence predominantly through antenatal clinic (ANC) sentinel surveillance (National AIDS Control Program, 2000). Surveillance is conducted every one to two years using a consistent methodology in the same population group, namely pregnant women who attend antenatal clinics (ANC).[2] Malawi's HIV sentinel surveillance indicates that HIV prevalence among antenatal attendees increased rapidly between the late 1980s and early 1990s. By the middle of the 1990s, prevalence stabilized and has since remained fairly constant.

ANC sentinel surveillance systems use unlinked, anonymous, methods for specimen collection and testing to avoid participation bias which can significantly affect HIV prevalence estimation. However, there remain many other potentially important biases inherent in sentinel data: ANC sites are often concentrated in more urban or readily accessible locations; pregnant women may be having unprotected sex at a greater rate than the general population of women; women with potential HIV-associated infertility are not captured; pregnant women that come to the ANC sites are likely to be more educated; and so on. In addition, extrapolations from pregnant women to the population as a whole (men and women) are often based on questionable assumptions.

It has been suggested that weighting of sentinel data can help overcome some of the above biases (Eckert et al, 2002). However, such corrections are likely to address, at best, only those biases arising from variables observed in the sentinel data, such as age, education, and place of residence. Moreover, they can only be applied to the age groups of women observed in the sentinel data. An additional problem with sentinel data encountered in many countries, including Malawi, is that not all districts, or other sub-national units, have sentinel sites.[3] As a result there are geographical pockets for which even rough estimates of prevalence are not available. Moreover, lack of broad coverage

[2] The system has collected data from 19 sentinel sites dating back to 1994. Some sentinel sites started data collection in 1990.
[3] In Malawi 8 out of 26 districts do *not* have sentinel sites.

implies that extrapolation to the national level, using the prevalence rates from sentinel data, may be very imprecise. Indeed, even in those districts that have sentinel sites the precision of the prevalence rates cannot be estimated as sentinel survey is not population based, and thus standard errors of the prevalence estimates cannot be estimated.

The only way to overcome the many biases inherent in sentinel data is to draw a nationally (and, ideally, sub-nationally) representative sample of the population and have respondents tested for HIV. National population-based surveys represent a much wider proportion of the population than do sentinel sites, since such surveys include non-pregnant women and also men, and they usually cover a large geographical area. Malawi is one of only a few countries where, for the first time in 2004, a Demographic and Health Survey (DHS) was fielded with the aim of collecting nationally representative data on the prevalence of HIV (NSO, Malawi and ORC Macro, 2005).[4] In Africa, such surveys still remain scarce as they are costly to implement. Moreover, even nationally representative household surveys are not immune from bias because of possibly significant non-response (Mishra et al., 2006). For instance, according to the 2004 DHS, in Malawi 30 percent of surveyed women and 37 percent of men refused to accept HIV testing. Considering the overall rate of participation of 96 percent for women and 86 percent for men, 33 percent of women in the original sampling frame, and 46 percent of men, were not tested for HIV. Nonparticipation in an HIV test due to absenteeism or refusal to be tested creates several associated biases (NSO, Malawi and ORC Macro, 2005). In order to account for the effects of non-response, HIV prevalence estimates could potentially be corrected using statistical procedures. Following application of such adjustments to the 2004 Malawi DHS data the estimate of HIV prevalence increased from 13.3 to 14.4 percent,[5] (NSO, Malawi and ORC Macro, 2005).[6]

[4] Results from the first eight national surveys implemented during 2001-04 in African counties are discussed in Mishra et al. (2006).

[5] The adjusted prevalence among women and men age 15-49 was 14.7% for women and 12.5% for men. There was a substantial difference in prevalence between urban and rural areas – 18.3% vs. 11.3%, respectively. Adult HIV prevalence, extrapolated from the 2001 ANC data, was estimated at 15% (25% in urban areas and 13% in rural areas). In line with the overall global trends, the HIV *incidence* rate in Malawi is estimated to have stabilized since late 1990s, and is estimated at about 1% of the overall population (Bello et al, 2006). HIV prevalence for women in the 2001 ANC data is reported in Table 1.

[6] Boerma, Ghys and Walker (2003) argue that estimates derived from such surveys are likely to be lower than true population prevalence, but the magnitude of the bias varies between countries.

When non-response bias can be satisfactorily accounted for, nationally representative surveys that test for HIV prevalence are clearly preferred among the available survey tools; they are generally expected to provide more accurate information on HIV prevalence than sentinel survey data. However, most countries are not yet in the position of being able to afford such surveys, or of repeating them frequently. At the same time, urgent calls for better and more timely response to the HIV epidemic raises demand for sub-national estimates of HIV prevalence. This demand arises from the recognition that the geographic distribution of HIV prevalence is likely to reflect "pockets" of high prevalence in some areas and low prevalence in others. At the aggregate level such heterogeneity may not be apparent, constraining policy makers from tailoring their responses to the specific geographic, and other, circumstances of different localities. There is thus heightened awareness of a need to better understand the correlates of HIV status, including the association between HIV and geographical characteristics. A question that arises in this context is whether it is possible to obtain reliable sub-national estimates of HIV prevalence using existing data sources, such as sentinel data coupled with other sources of information.

This paper examines the case of Malawi, where the adult prevalence is estimated to be the 8th highest in the World (UNAIDS, 2006), and demonstrates how one can utilize small area estimation (SAE) methods to obtain a sub-national (district) level profile of HIV prevalence in a setting where national population-based surveys of prevalence (such as the 2004 DHS) are absent. The approach considered here draws from the small area estimation approach applied to welfare measurement (Elbers et al., 2003). It represents the first time that small area estimation methods have been applied to HIV prevalence data.

The methodology presented in this paper combines HIV status information available from sentinel surveillance data (the 2001 ANC sentinel survey) with representative population-based data from a household survey (DHS 2000) as well as Census and GIS data. The proposed method essentially involves 3 stages. In the first stage information on individual characteristics from the DHS 2000 (such as age and education of women aged 15-49) is used to weigh the respective individual information from the sentinel survey. This exercise is carried out in order to render the sentinel data as representative of the underlying population as possible. In the second stage, the

resulting weights are used while drawing on the sentinel data to estimate a probit model of the association between HIV status and a set of individual, household and locational characteristics (with some of the latter coming from the census and other data linked-in using GIS). There is no attempt to estimate a causal model; rather the objective is to uncover the conditional correlation between HIV status and a set of "explanatory" variables that happen to have been collected in both the sentinel data as well as the household survey. Third, HIV status is imputed into the 2000 DHS household survey on the basis of parameter estimates from the probit model applied to identically defined individual, household and locational characteristics in the sentinel survey. HIV prevalence for each target woman in the DHS sample is estimated,[7] and then aggregation is performed to get the estimate of prevalence at the desired level (e.g., district).[8]

In our example we predict HIV prevalence into the nationally representative DHS survey for 2000 that does not separately collect prevalence data. However, as mentioned above, in 2004 another DHS survey was fielded in Malawi that did collect prevalence data. In an effort to check the validity of our imputation procedure we compare district-level predicted prevalence in the 2000 DHS with observed prevalence at the district level from the 2004.[9] Because there is a general perception that prevalence rates changed little during this time period we look to see whether the district level estimates we predict in the DHS 2000 are reasonably close to those that are directly estimated from the 2004 DHS survey. [10],[11]

[7] In essence, this is an out-of-sample prediction.

[8] Case studies using similar approach to predict welfare estimates show that the method provides unbiased estimates with relatively small standard errors, and are precise enough to allow for comparisons across geographic areas (Hentschel et al., 2000).

[9] We anchor our estimates to the 2000 DHS because other data used (sentinel, census, GIS) have also been collected around that time. More discussion on this is provided later in the paper.

[10] This validation using the 2004 DHS-based prevalence assumes that prevalence has not changed much between 2001 and 2004. Malawi National AIDS Commission and UNAIDS estimates indicate that indeed the prevalence rate in Malawi was stable during that period of time (National AIDS Commission, 2003; UNAIDS, 2006).

[11] Elbers et al (2003), use bootstrap simulations to calculate the prediction error of the district-level estimate when estimating poverty rates after combining household survey data with population census data. This prediction error consists of three components: idiosyncratic error, which is the deviation of the actually observed prevalence from its expected value; model error, which is due to the variance in the parameter estimates; and computation error, which arises from the way the expected value of the prevalence is computed. We produce standard error calculations based on the same procedure, but these are partial only, as they do not reflect the additional contribution to the overall standard errors stemming from the fact that the DHS 2000 survey is a sample survey rather than population census.

The remainder of the paper is organized as follows. Section 2 presents the methodology and data used for estimating the district-level prevalence in the absence of nationally-representative population surveys. Section 3 presents the regression results and discusses how the district-level estimates of HIV prevalence obtained from the application of the SAE methodology to the DHS 2000 survey compare with the sentinel data estimates for 2001 and the "gold-standard" prevalence rates derived directly from the 2004 DHS Survey. Section 4 concludes by reviewing the main findings.

2. Methodology and Data

2a. The basic idea

The methodology used in this paper draws on the recently developed small area estimation techniques applied to welfare measurement (Elbers et al., 2003).[12] The idea is straightforward. Let p_i be an indicator of individual HIV infection found in the sentinel data. We first model p_i as a function of individual level variables (such as age and education) found in the 2001 sentinel (ANC) survey, and of selected commune/district means derived from the 1998 Malawi National Census data (Benson, 2002), the 2001 Health Facilities data, 2001 GIS data, and the 2000 Demographic and Health Survey (DHS) data.[13] Such commune/district means are useful because individual infection status is clearly seen as dependent not only on individual characteristics, but also on regional factors. The construction of regional means is driven by a priori expectations on which regional factors are likely to be correlated with the probability of HIV infection, as well as by data availability. Importantly, the set of explanatory variables is restricted to those that can be linked to individuals in the DHS 2000 sample. Second, the set of parameter estimates resulting from the estimation of the model is applied to the identically defined variables in the nationally representative DHS 2000 sample (again,

[12] The important distinction is that while in the welfare measurement the household survey is representative of the total population of households, in the prevalence measurement ANC data are representative of only the respective sub-population of all population of women. The paper discusses in a greater detail below the possible ways to deal with this caveat.

[13] We potentially could use the 2004 DHS data instead of the 2000 DHS data, but since the reliability of estimates depends on the various data sources been closer to each other in time, we use the 2000 DHS which is more consistent with the 2001 ANC survey, 2001 GIS data, and the 1998 Census.

these variables will consist of the individual level characteristics such as age and education that are identically defined in both the DHS2000 and the sentinel dataset, and the same commune/district means that were inserted into the sentinel dataset for the probit model) to obtain the predicted probability of being HIV/AIDS positive for each woman in the DHS2000. Third, individual-level HIV status indicators are aggregated up to our desired geographic level to obtain the estimate of HIV/AIDS prevalence for respective sub-population groups. The level of geographic aggregation is restricted to the level at which DHS2000 data is nationally representative, which is a district unit in our case.[14] The section below describes in a greater detail each step involved in the analysis.

2b. The prediction model

The first concern is to develop an accurate empirical model of p_i, the individual HIV status observed in 2001 sample of women attending ANC sites. Since the dependent variable is a dichotomous one, the estimation is performed using maximum likelihood probit. The specification of the model in general terms can be presented as:

$$\Pr (p_i = 1 \mid \mathbf{X}_i, \mathbf{X}_r) = \Phi (\mathbf{X}_i\alpha + \mathbf{X}_r\beta) \quad (1)$$

where $\mathbf{X}_i$ is a vector of individual level characteristics, and $\mathbf{X}_r$ is a vector of district/sub-district characteristics affecting a person's chance of being HIV infected, α and β are the vectors of parameter estimates. Φ denotes the standard normal cumulative distribution function.

The choice of $\mathbf{X}_i$ variables is restricted to those individual characteristics observed in the 2001 sentinel survey data which also occur and are identically defined in the DHS 2000 data. These variables are individual's age and educational status. Since the age distribution for women in the DHS sample is from 15 to 49 years, we also restrict the

[14] Note that using unit record census data as a base for prediction would allow obtaining the accurate estimates of HIV prevalence for much smaller administrative units. However, we do not have access to such data. It is also important to bear in mind that aggregation does not need to be geographical. One can, for example, estimate prevalence by age groups, etc.

sentinel survey sample to women of the same age group.[15] The application of the methodology is based on the assumption that the age/gender distribution for women in the DHS closely reflects that observed in the sentinel data.[16] However, this is not likely to be the case since the sentinel survey refers to pregnant women who attended sentinel sites, while DHS covers all women age 15 to 49. Two important issues that need to be investigated in this context, before we can confidently move ahead, are thus whether: 1) any sub-group of women in the DHS sample matches sentinel survey women in terms of age/education distribution, so that we could restrict our predictions to that specific sub-group; and 2) to what extent the population of women residing in districts with sentinel sites is representative of women nationwide?

The first issue is important because if the total DHS sample is far from being close to the sentinel data by age/education distribution, we have to focus on the sub-group of DHS women closest to sentinel survey women (such as pregnant women or antenatal clinic attendees) and/or we must reweigh the sentinel survey data so that it becomes representative by age/education of a particular sub-group or all women in the DHS.

The second issue is important because a crucial assumption we make in estimating the model based on sentinel survey data is that no strong selectivity exists in terms of sentinel sites' placement across the districts (i.e., the population of women from districts with sentinel sites is fairly representative of the total population of women).[17]

The first issue is investigated by looking at the age/education distributions for various sub-groups of women in the DHS 2000 (all women, ANC attendees who gave birth last year, ANC attendees that agreed to having a blood sample drawn and who gave a birth last year, currently pregnant women) and comparing them to the distributions for

[15] This means that we have to drop only 15 out of 7300 observations for which age is observed (the total sample size of 2001 sentinel survey is 7372).

[16] The key issue in extrapolating from antenatal clinic data to a population of all women (age 15 to 49) is that in the ANC data there is literally no information on women who have not recently been pregnant. Hence, similar to other matching techniques, such as the Propensity Score Matching, the proposed method assumes that these women do not differ significantly on individual *unobserved* variables from those who do get captured. This assumption may be reasonable given that the average fertility rate in Malawi is 5.7 children born/woman. To check for the robustness of the results, predictions are made not only for all women (age 15-49) in the DHS, but also for the sub-samples of the currently pregnant women and ANC attendees in the DHS. Those results are reported and discussed further in the paper. One should also keep in mind that the assumptions made when extrapolating the prevalence directly from the ANC data to the total population (as it is done conventionally in many countries) are more involved and less plausible compared to those that are involved in the SAE methodology.

[17] Note that only 18 out of 26 districts have sentinel sites.

those women in the sentinel survey data. The second issue is investigated by comparing, for each sub-group of women, distributions for women from sentinel site districts with those for all women. The results are presented in Table 2.

We can draw several conclusions based on these results. First, the sample of all women in the DHS 2000 is clearly different in its age/education structure from all other sub-samples. Second, the sample of currently pregnant women in the DHS 2000 is very similar in terms of age and education distributions to the sample of the last year ANC attendees in the DHS 2000. The only exception is that the share of the youngest group (age 15 to 19) is higher for currently pregnant women (probably because the youngest are less likely to seek antenatal care). Third, the only difference between all ANC attendees and those ANC attendees who gave the blood sample is that the latter tend to be somewhat more educated. Fourth, the sentinel survey sample seems to be closest in terms of age and education distribution to the samples of currently pregnant women and ANC attendees in the DHS. Despite this broad similarity it is important to note that the sentinel survey sample is younger and more educated (it has similar share of women with primary education, but a noticeably higher share of women with secondary+ education).

The findings outlined above indicate that basing our predictions in the DHS2000 survey either on the sample of last year ANC attendees or on currently pregnant women should not lead to significant differences in results. However, the former sub-group is probably preferable due to its larger size, which is especially important considering that our predictions will be at the district level. In addition, since even the sample of ANC attendees is somewhat different by age/education distribution from the sentinel survey sample, we will need to re-weight the sentinel data to correct for this problem, i.e. to make both distributions match by age/education. It is noteworthy that comparing the distributions by age/education for districts with sentinel sites to the distributions for all districts we find no statistical difference between the samples (for all groups of women), indicating that the population of districts with sentinel sites reflects well the total population of women (Table 2).

The next step involves the calculation of weights that make the distribution by age/education in the sentinel survey reflect that of a specific group in the DHS. We construct two sets of weights. The first set forces the age and education distribution in the sentinel survey to mirror that of ANC attendees while with the second set the age and

education distribution of the sentinel survey mirrors that of all women in the DHS. Note that for age/education cells with less than 20 observations we do not reweigh the data (i.e., the weight of unity is assigned). The first/second set of weights will be used in the probit models from which we will then make DHS based predictions for ANC attendees and for all women (aged 15-49).[18] The age/gender distributions for various groups of women, and the resulting sets of weights, are presented in Tables 3a-3b.[19] The results indicate that for any given age group, the sample of women from the sentinel survey has a higher share of those with secondary+ education. The constructed weights correct for this problem.

The next step of the analysis involves constructing a set of variables at the commune/district level (vector $\mathbf{X}_r$ in our model) that are likely to influence an individual's chance of HIV infection. To do that we draw variables from a rich set of data sources including the 1998 Malawi National Census, 2001 Health Facilities data, 2001 GIS data, and the 2000 DHS. Table A1 in the Appendix presents the list of district/sub-district level variables used in the analysis, their sources, units of variables' measurement, initial level of aggregation, the level of aggregation used in the analysis, and the rationale for these variables. To use these commune/district level means in our analysis we merge both the sentinel survey and the DHS data with these regional variables on the basis of the identification numbers for respective geographic units (EA/TA/district).

Obviously, not all of the variables listed in Table A1 can enter the regression due to a high degree of collinearity among many variables. Hence, the subsequent step of the analysis involves selecting a sub-set from all possible regional variables to enter the final regression model. A separate model is estimated for each of the three major regions in the country (North, Central, and South) to allow parameter estimates to vary across regions. For each region we select a specification of the model that achieves a high degree of

[18] Note that although by re-weighing the sentinel data we can correct for the differences in age/gender distributions across the two data sources (sentinel survey and the DHS 2000), there are likely to be the differences in other unobserved characteristics that we are not able to control for. Due to this reason the sample of ANC attendees in the DHS, which is initially (before re-weighing) closer to the sentinel survey sample in terms of its age/education structure, is likely to give more accurate predictions (since it is likely to have fewer differences in unobservable characteristics as well).

[19] Further discussion in this paper focuses on the results for all women aged 15-49 for the purpose of comparison with the same age group in the sentinel data and the DHS.

predictive power. As mentioned above, when estimating the model we use a set of weights that make sentinel survey reflect a specific group of women in the DHS in terms of age/education distribution. We also take into account the clustering effect which results from the fact that there are groups of observations which come from the same sentinel site, and thus there are some unobservable characteristics common for all women from a specific sentinel site. The empirical results will be discussed later in the paper. We next describe how we obtain district-level predictions (point estimates).

2c. Predictions of HIV prevalence

The estimation of regression equation (1) using the sentinel data (merged with the district/sub-district means) produces the vectors of parameter estimates α and β, as well as the vector of residuals. Note that since a separate specification is estimated for each region, these vectors will be region-specific.

The predicted values of HIV prevalence are computed in the following way. We first obtain the empirical distributions of α and β from the model by making 100 draws using bootstrap simulations. We denote these vectors as α_n and β_n, where n denotes the number of rows in the vector (with the number of columns being equal to the number of variables in the model). Next, to obtain the vector of simulated errors we save the vector of residuals after the model estimation, and then make 100 random draws with replacement from this vector using bootstrap simulations. We denote this drawn vector of residuals as $\mathbf{e_n}$ (with the number of columns being equal to one). The vectors α_n, β_n, and $\mathbf{e_n}$ are then added to form a matrix, which we can name $\mathbf{B}$.[20] Each row of values found in this matrix is applied to the values of respective variables in the DHS 2000 data to make a prediction of individual's probability of being infected for each observation in the DHS. This can be expressed as:

$$\Pr^{in}(p_i = 1 \mid \mathbf{X}_i, \mathbf{X}_r, \mathbf{e_n}) = \Phi(\mathbf{X}_i \alpha_{\mathbf{n}}^{n} + \mathbf{X}_r \beta_{\mathbf{n}}^{n} + \mathbf{e_n}^{n})$$

[20] Note that since a separate specification is estimated for each region, the matrix **B** will be region-specific as well.

where superscript *i* refers to a given observation, and superscript *n* refers to the number of the row in the B matrix used in the estimation. As a result, for each observation (individual) we obtain the number of predictions equal to *n*.

To obtain the estimate of prevalence at the district level (or any other level of aggregation we are interested in) we first calculate the mean value by district for each column of observations, denoted $\Pr^n$. Note, the calculation of the means takes into account the sample frame of the DHS sample (i.e., stratification and clustering). As a result, for each district we obtain the vector of *n* such means (with *n*=100 in our case). The mean calculated over those *n* means will be our point estimate of predicted prevalence.[21]

3. Empirical Results

3a. Regression results

We first report and discuss the regression results which form the basis for predicting the HIV prevalence into the DHS 2000 and estimating prevalence rates at the district level. Table 4 presents the results based on the set of weights that make the distribution by age/gender for women in the sentinel survey reflect that for all women in the DHS. As mentioned earlier the estimated coefficients represent correlations, not impact parameters. Omitted variable bias is thus likely to afflict these models and as such the interpretation of estimated parameters is not terribly meaningful. Nonetheless, we offer a few comments on the estimated correlations based on the results reported in Table 4.

[21] As mentioned earlier, a sense of the precision of our prevalence estimate can be obtained by calculating the standard deviation of those *n* means calculated above. This statistic will capture two important components of the overall total prediction error: *idiosyncratic error* and *model error*. A further component, *computation error* is attributable to the employment of a bootstrap simulation approach. With the number of simulations (*n*) equal to 100, the computation error is equal to $\sigma/10$, where σ is the standard deviation of the distribution of means. Thus by choosing a large enough number of simulations, the computation error can be set arbitrarily small. By combining the idiosyncratic, model and computational errors we obtain the prediction error. To this error one would still need to add the sampling error attributable to the DHS 2000 sampling design in order to obtain the final, total, predication error (see Elbers et al, 2003, and Elbers et al 2008, for further discussion).

Urban/major city location is associated with a higher risk of HIV infection – this finding is consistent across regions. Mean population age and share of women who report knowledge of HIV prevention practices appear significant only in the Central region.[22] We find a positive correlation between the share of people living below the poverty line and the probability of being HIV-positive in the North and South regions.

The regression results for all districts show an inverted U-shape relationship between the HIV infection status and age – the probability of being HIV-infected first increases to about age of 30, and then decreases. This is consistent with the "benchmark" 2004 Malawi DHS data which show that HIV prevalence peaks for women and men age 30-34 (NSO, Malawi and ORC Macro, 2005).

Education level is found to be correlated positively with the risk of being HIV-infected in the North and Central regions. This correlation is also validated by the 2004 Malawi DHS data, which show higher HIV prevalence among women with secondary or higher education. Interestingly, the 2004 Malawi DHS showed no significant differences in the testing refusal rates among women with various education levels.

The share of orphans in the TA is positively correlated with positive HIV status in the regression specifications for all regions of Malawi. That is not surprising given that higher HIV prevalence in a sub-district results in the larger death toll and hence in the higher share of orphans.

There is a negative correlation between the mean age at first sex and the mean age at first marriage (for women) in the TA and the individual probability of being HIV-positive. We also find a negative correlation between the share of women in TA who report knowledge of HIV prevention methods and the probability of being infected in the Central region.

All in all the results described above appear broadly consistent with our expectations and with prior work on HIV in Malawi.

[22] Note that if the knowledge of HIV prevention practices is almost universal across TAs in the region, this variable is not expected to be statistically significant.

3b. Predicted prevalence at the district level

Based on the regression results discussed above we predict HIV prevalence for women aged 15-49 in the DHS 2000 survey. The actually observed district-level prevalence rates (in the 2001 sentinel data), predicted prevalence rates based on the 2001 sentinel survey (in-sample), and SAE-based predicted prevalence rates in the 2000 DHS survey (out-of-sample) are shown in Table 5. The reported estimated (partial) standard errors provide a sense of the accuracy of our SAE based district-level predictions. Note, however, that these standard errors are understatements of the true standard errors (in the same way that the sentinel-data based estimates are also subject to some kind of sampling error).

We first notice that although the regressions are estimated at the regional (North, Central, South) level, the estimates of HIV prevalence based on in-sample predictions into the 2001 sentinel survey data are generally quite close to the observed prevalence.[23] This indicates that the regression model estimated at the region level is able to provide a good fit at the district level as well.

The out-of-sample predictions (those that use 2000 DHS sample) indicate lower prevalence than sentinel survey estimates. The sentinel survey actual prevalence at the national level of 19.4% drops to 13.2% for the 2000 DHS-based predictions. The lower predicted prevalence for the 2000 DHS sample is to be expected given the selective nature of the sentinel sample.

It is noteworthy that predicted HIV prevalence rates in the DHS sample at the regional level mask substantial heterogeneity across districts – even for the districts belonging to the same region. For instance, in the South region the estimated prevalence varies from 11% for Mwanza district to 25.6% for Mulanje (Table 5).[24]

How good are our predictions using the small area estimation (SAE) methodology compared to our "gold standard" of HIV prevalence rates derived directly from the 2004 Malawi DHS? Table 6 reports estimates of HIV prevalence derived directly from the

[23] With the exception of Chiradzulu, Machinga, Mangochi, and Mulanje districts in the South.
[24] One would generally expect the prevalence for the sample of all women age 15-49 to be generally lower than the prevalence for the sample of ANC attendees or currently pregnant women age 15-49. We find that this is indeed the case for most districts in Malawi. Those results are available from authors upon request.

2004 DHS for a selection of 9 districts, as well as estimates at the all-Malawi level.[25] These can be compared to predicted HIV prevalence estimates in the DHS 2000. Assuming fairly stable prevalence rates over this time period, a comparison of 2004 district level estimates versus those for 2000 can provide an informal validation check on the small area estimation procedure outlined in this paper. At the national level, the SAE methodology yields an estimated prevalence of 13.2% in the DHS 2000, which is very close to the prevalence of 13.3% derived directly from the 2004 DHS (Figure 1, and Table 6). When we compare the predictions across the North, Central and South regions, we find, again, that the predicted prevalence rates in the DHS 2000 are quite close to those obtained directly in the DHS 2004, while being quite distinct from those obtained directly from the 2001 Sentinel survey (Figure 1).

District level estimates are, on the whole, also fairly close (Table 6). For example in Blantyre HIV prevalence amongst women aged 15-49, based on the SAE procedure applied to the 2000 DHS, is estimated at 20.4%. This can be compared to 22.5% observed directly in the 2004 DHS. Similarly, in Mzimba the 2004 DHS yields an estimate of 6.4%, only a bit lower than the SAE based estimate of 8.6% in the 2000 DHS.

While these results are encouraging, it is important to note that agreement between the 2004 DHS and SAE-based predictions from the 2000 DHS is not inevitable. Table 7 illustrates the case for the district of Lilongwe. Observed prevalence in the 2001 Sentinel survey is estimated at 20.0%. This is markedly higher than the predicted prevalence, of 11.1% based on the SAE methodology and the DHS 2000. It is even more dramatically at odds with the prevalence estimate observed in the 2004 DHS of 1.6%. NSO (2005) discusses serious concerns with non-response in Lilongwe and presents an alternative prevalence estimate for the 2004 DHS, after adjustment for non-response, of 11.5%. This adjusted estimate is very close to the SAE predicted prevalence estimate but remains a good deal lower than the sentinel survey based estimate. A second example of disagreement between the 2004 DHS estimate and the SAE based estimate from the 2000 DHS concerns the district of Zomba (Table 6). In this case the 2004 DHS estimate of 24.6% is markedly higher than the SAE estimate of 13.4% based on the 2000 DHS. It is unclear why such glaring disagreement should arise in this case, but again it is

[25] We draw on National Statistical Office (NSO) of Malawi, 2005, for the 2004 DHS-based estimates.

conceivable that response biases are at play. NSO (2005) reports, for example, that the 2004 DHS data point to an HIV prevalence in Zomba amongst men aged 15-49 that is roughly on par with the national average for men of that age (10.5% in Zomba versus 10.2% in Malawi as a whole). For reasons not addressed in the MNSO report, the 2004 DHS data suggest that prevalence amongst women in Zomba (24.6%) is nearly twice the national figure for this population group (13.3%). The SAE-based estimate, on the other hand, suggests that prevalence amongst women in this district is also on par with the national figure for women (13.4% versus 13.2%). These examples serve to remind that while the SAE methodology is certainly likely to introduce a degree of uncertainty into assessments of HIV prevalence, it is also the case that "gold-standard" estimates, based on direct measurements, deserve a degree of circumspection.

4. Conclusion

In most developing countries, including Malawi, estimates of HIV prevalence are obtained through sentinel surveys of women visiting antenatal clinics (ANCs). Sentinel surveys are often the only source of information on HIV prevalence in a country, and as such are extremely useful. However, they are subject to numerous limitations. These surveys are not based upon a probability sample and are not representative of the population as a whole. Moreover, they do not provide reliable information on prevalence at the sub-national level.

In recent years, several countries in Africa have implemented Demographic and Health Surveys (DHS) that include HIV testing of adult population.[26] These surveys are generally large in size and thus usually permit the calculation of prevalence rates at some sub-national level. Although these surveys are known to suffer from non-response bias, they unquestionably represent an improvement over sentinel surveys. However, such population-based surveys of prevalence rates are still very rare. There is thus an interest to search for improved methods for small-area estimation (SAE) of HIV prevalence which would make it possible to produce more accurate (compared to those extrapolated directly from the sentinel surveys) sub-national estimates of HIV prevalence when population-based surveys of prevalence are not available.

This paper proposes a new approach to the estimation of HIV prevalence for relatively small geographic areas. The proposed approach is believed to overcome some of the difficulties/biases involved in currently existing methods of deriving HIV prevalence (at both national and sub-national levels) directly from sentinel surveys. The paper provides justifications why a new method could be useful, and also describes the limitations of the proposed approach.

The methodology proposed in this paper is similar in spirit to the poverty mapping methodology introduced by Elbers et al.(2003)**.** The method essentially involves 3 stages. In the first stage the information from a DHS (or similar survey) on individual characteristics (such as age and education) of women (age 15-49) is used to weight the respective individual information from the sentinel antenatal clinics. In the second stage,

[26] The full list of those countries is provided in the UNAIDS Report (UNAIDS, 2006).

the resulting weights are used in estimating a probit model with the sentinel data that includes individual information combined with the geographical/administrative variables (some of the latter coming from the census and other data linked in using GIS) to obtain a set of parameter estimates on correlates of HIV status (with no claim to causality). In the third stage, these parameter estimates are applied to the identically defined variables in the DHS (or similar) survey to predict HIV prevalence for each target woman in the DHS sample. Aggregation is then performed to obtain an estimate of prevalence at the desired geographic level.

The estimates produced using this SAE approach are then compared to those derived directly from the sentinel data in Malawi. The SAE estimates suggest lower HIV prevalence than is found in sentinel data. We further compare SAE estimates with "gold-standard" estimates of HIV prevalence obtained directly from the 2004 Malawi DHS. At the national level, the SAE methodology yields an estimated prevalence of 13.2% in the 2000 DHS. This is very close to the 13.3% estimate obtained directly from the 2004 DHS. We note that there are good reasons to expect little change in HIV prevalence between 2000 and 2004 and so we regard this close agreement between SAE estimate for 2000 and observed prevalence in 2004 as indicating that the SAE method is yielding sensible estimates. Close agreement between the 2000 SAE estimates and 2004 DHS estimates is also found at the regional level and, to a considerable extent, also to the district level in Malawi. An important area for future research is to understand better the circumstances underlying those cases where district level predicted prevalence rates in the DHS 2000 disagree with 2004 district estimates. There are many factors that could underlie this disagreement, and it is important to acknowledge that they could stem also from non-response problems in the 2004 data.

In sum, the results generally indicate that the SAE methodology introduced here can potentially hold a lot of promise. First, it produces estimates of prevalence that are lower then those derived directly from sentinel data. This is good news because in all countries where recent DHS surveys have included direct HIV testing the findings confirm that sentinel data overestimates prevalence (UNAIDS, 2007). Second, the SAE methodology produces estimates of prevalence which are quite close to the "gold-standard" prevalence rates obtained from a nationally representative survey that directly collects information on HIV prevalence. Third, given the right set of data (and

assumptions) the SAE methodology can potentially produce estimates of prevalence at a more disaggregated geographical level than the DHS.[27] As argued above, it is worth noting that the "gold standard" of population-based surveys testing for HIV is also subject to error, considering that significant adjustments to the survey data might still be needed to account for non-response. Hence, the SAE methodology could potentially be used as a tool to verify DHS-based results.

In closing it is important to emphasize that the SAE methodology outlined here should be further tested in countries where direct testing for HIV was undertaken in order to provide further verification of the approach. Further work is also needed to develop more complete estimates of the precision of predicted prevalence rates. A potential area of further research might also be to explore how HIV prevalence obtained via this approach can predict HIV *incidence* compared to conventional estimates from surveillance and DHS surveys.

[27] For example, if one makes the predictions based on the census sample.

References:

Bello, G. A., J. Chipeta and J. Aberle-Grasse. 2006. Assessment of trends in biological and behavioral surveillance data: is there any evidence of declining HIV prevalence or incidence in Malawi? Sexually Transmitted Infections 2006;82(1): i9-i13.

Benson, T. 2002. Malawi: An Atlas of Social Statistics.

Boerma, J. Ties, Ghys D. Peter, Walker Neff. 2003. Estimates of HIV-1 Prevalence from National Population-based Surveys as a New Gold Standard. The Lancet 363, 1929-30.

Crampin A.C., Glynn J.R., Ngwira B.M.M., Mwaungulu F.D., Ponnighaus J.P., Warndorff D.K. and Fine P.E.M. 2003. Trends and measurement of HIV prevalence in northern Malawi. AIDS 17(12), 1817-25.

Eckert E. L., Damisoni H., Bicego G., Martin R., 2002. A Validation of Malawi's Sentinel Surveillance Data. Memo. XIV International Conference on AIDS, Barcelona, July 7-12.

Elbers, C., Lanjouw J. O., Lanjouw P., 2003. Micro-Level Estimation of Poverty and Inequality. Econometrica 71(1), 355-64.

Elbers, C., Lanjouw, P., and Leite, P. G., 2008. Brazil within Brazil: Testing the Poverty Mapping Methodology in Minas Gerais. Policy Research Working Paper 4513, the World Bank.

Garbus, L. 2003. HIV/AIDS in Malawi. Policy Research Paper. AIDS Policy Research Center, University of California San Francisco.

Hentschel, J., Lanjouw, J.O., Lanjouw, P., and Poggi, J. 2000. Combining Census and Survey Data to Trace the Spatial Dimensions of Poverty: A Case Study of Ecuador. World Bank Economic Review 14(1), 147-65.

Mishra, V., Vaessen M., Boerma J. T., Arnold F., Way A., Barrere B., Cross A., Hong R., and Sangha J. 2006. HIV Testing in National Population-based Surveys: Experience from the Demographic and Health Surveys.

National AIDS Commission of Malawi. 2003.http://www.policyproject.com/pubs/countryreports/MALNatEst2003.doc Estimating National HIV Prevalence in Malawi from Sentinel Surveillance Data: Technical Report.

National Statistical Office (NSO) of Malawi and ORC Macro. 2005. Malawi Demographic and Health Survey, 2004. Calverton, Maryland: NSO and ORC Macro.

UNAIDS. 2006. Report on the Global AIDS Epidemic. Geneva.

UNDP. 2003. Zimbabwe Human Development Report 2003: Redirecting our responses to HIV/AIDS.

Tables

Table 1: HIV/AIDS prevalence for women in Malawi as pictured by the 2001 sentinel (ANC) data

Variable	N of obs.	Prevalence, %
region		
North	2,006	15.95
Central	2,633	17.55
South	2,719	24.05
education		
none	1,326	21.72
primary	4,630	17.60
secondary+	1,281	23.73
unknown, missing	121	23.97
age group		
<19	1,572	11.83
20-24	2,759	20.19
25-29	1,643	24.59
30-34	837	22.10
35+	475	18.95
unknown, missing	72	19.44
Total	7,358	19.52

Table 2: Comparing the age/education distributions in the 2000 DHS and 2001 sentinel (ANC) survey data for various groups of women

	All women (2000 DHS)				ANC attendees in last year (2000 DHS)				ANC attendees in last year who gave blood sample (2000 DHS)				Currently pregnant women (2000 DHS)						**ANC attendees (2001 Sentinel Survey)**
	all districts			only districts with sentinel sites	all districts			only districts with sentinel sites	all districts			only districts with sentinel sites	all districts			only districts with sentinel sites			
	mean	95% confidence interval		mean	mean	95% confidence interval		mean	mean	95% confidence interval		mean	mean	95% confidence interval		mean	95% confidence interval		mean
Age																			
15-19	21.69	*20.98*	*22.39*	21.56	14.87	*13.51*	*16.22*	14.69	14.02	*11.89*	*16.14*	13.25	18.41	*16.48*	*20.34*	18.49	*16.26*	*20.72*	21.43
20-24	22.37	*21.66*	*23.08*	22.63	33.54	*31.75*	*35.34*	33.88	33.44	*30.55*	*36.33*	33.62	32.86	*30.52*	*35.19*	32.30	*29.61*	*34.99*	37.90
25-29	18.16	*17.50*	*18.82*	18.45	25.36	*23.70*	*27.01*	25.70	28.22	*25.46*	*30.97*	28.95	22.39	*20.32*	*24.46*	23.44	*21.00*	*25.87*	22.64
30-34	11.85	*11.29*	*12.40*	12.00	12.35	*11.10*	*13.60*	11.97	10.95	*9.04*	*12.87*	11.74	13.34	*11.65*	*15.03*	13.60	*11.63*	*15.57*	11.53
35-39	10.77	*10.24*	*11.30*	10.76	8.99	*7.90*	*10.08*	8.86	8.99	*7.24*	*10.74*	8.21	7.91	*6.56*	*9.25*	7.27	*5.78*	*8.76*	5.15
40-44	7.97	*7.51*	*8.43*	7.74	3.30	*2.62*	*3.98*	3.17	2.63	*1.65*	*3.60*	2.29	3.49	*2.58*	*4.41*	3.44	*2.40*	*4.49*	1.28
45-49	7.19	*6.75*	*7.63*	6.85	1.59	*1.12*	*2.07*	1.74	1.76	*0.96*	*2.57*	1.94	1.60	*0.98*	*2.23*	1.47	*0.78*	*2.16*	0.08
Education																			
no education	27.04	*26.28*	*27.79*	25.30	28.44	*26.73*	*30.16*	27.28	24.12	*21.50*	*26.74*	22.67	27.14	*24.93*	*29.36*	26.37	*23.84*	*28.90*	18.44
primary	61.86	*61.03*	*62.69*	62.34	63.65	*61.82*	*65.48*	64.06	66.24	*63.35*	*69.14*	67.05	65.81	*63.46*	*68.17*	66.57	*63.86*	*69.28*	63.81
secondary+	11.11	*10.57*	*11.64*	12.37	7.91	*6.88*	*8.94*	8.65	9.64	*7.83*	*11.44*	10.28	7.04	*5.77*	*8.31*	7.06	*5.58*	*8.53*	17.75
Number of observations	13220	13220	13220	10239	2664	2664	2664	2073	1028	1028	1028	784	1557	1557	1557	1166	1166	1166	7285

Note: all tabulations using DHS are weighed;

ANC attenders (DHS) are women who had a birth in the last year and who had a prenatal care from a professional source (doctor/nurse/midwife);

Table 3a: The distribution of ANC attendees in the 2000 DHS vs. women in the 2001 sentinel survey by age group/gender, %

	DHS				*Sentinel Survey*				*weight*		
	Education level				**Education level**				**Education level**		
Age group	no education	primary	secondary+	Total	no education	primary	secondary+	Total	no education	primary	secondary+
15-19	2.18	11.52	1.17	14.87	1.91	15.71	3.77	21.39	1.14	0.73	0.31
20-24	6.60	22.82	4.12	33.54	5.16	23.88	8.93	37.97	1.28	0.96	0.46
25-29	7.89	15.37	2.10	25.36	4.80	14.26	3.53	22.59	1.64	1.08	0.59
30-34	4.74	7.29	0.32	12.35	3.59	6.78	1.21	11.58	1.32	1.08	0.26
35-39	4.24	4.54	0.21	8.99	2.18	2.65	0.29	5.12	1.94	1.71	0.72
40-44	1.81	1.49	0.00	3.30	0.77	0.49	0.01	1.27	2.35	3.04	1.00
45-49	0.98	0.62	0.00	1.59	0.04	0.04	0.00	0.08	1.00	1.00	1.00
Total	28.44	63.65	7.91	100.00	18.44	63.81	17.75	100.00			

Note: The sample sizes are: 2664 (2000 DHS); 7285 (2001 sentinel survey); only women age 15 to 49

ANC attenders (DHS) are women who had a birth in the last year and who had a prenatal care from a professional source (doctor/nurse/midwife);

All tabulations using DHS data are weighted

For cells with less than 20 observations we do not reweight the data (i.e., the weight of unity is assigned)

Table 3b: The distribution of all women in the 2000 DHS vs. women in the 2001 sentinel survey by age group/gender, %

	DHS				*Sentinel Survey*				*weight*		
	Education level				**Education level**				**Education level**		
Age group	no education	primary	secondary+	Total	no education	primary	secondary+	Total	no education	primary	secondary+
15-19	1.66	16.94	3.09	21.69	1.91	15.71	3.77	21.39	0.87	1.08	0.82
20-24	4.15	14.23	3.99	22.37	5.16	23.88	8.93	37.97	0.80	0.60	0.45
25-29	5.45	10.85	1.86	18.16	4.80	14.26	3.53	22.59	1.14	0.76	0.53
30-34	4.13	6.89	0.82	11.85	3.59	6.78	1.21	11.58	1.15	1.02	0.68
35-39	4.33	5.76	0.69	10.77	2.18	2.65	0.29	5.12	1.99	2.17	2.38
40-44	3.75	3.82	0.40	7.97	0.77	0.49	0.01	1.27	4.87	7.80	1.00
45-49	3.57	3.36	0.26	7.19	0.04	0.04	0.00	0.08	1.00	1.00	1.00
Total	27.04	61.86	11.11	100.00	18.44	63.81	17.75	100.00			

Note: The sample sizes are: 13220 (2000 DHS); 7285 (2001 sentinel survey); only women age 15 to 49

All tabulations using DHS data are weighted

For cells with less than 20 observations we do not reweight the data (i.e., the weight of unity is assigned)

Table 4: Probit regression results on the determinants of HIV infection status by region

(weights that make distribution by age/gender for women in the sentinel survey reflect that for all women in the DHS)

	North				Central				South			
Variable	dF/dx	Std. Err.	P>|z|	x-bar	dF/dx	Std. Err.	P>|z|	x-bar	dF/dx	Std. Err.	P>|z|	x-bar
Urban (city) location	0.018	0.003	0.000	0.804	0.094	0.027	0.003	0.816				
Blantyre (largest city)									0.368	0.079	0.000	0.195
Blantyre X Age									-0.004	0.001	0.000	4.859
Age	0.064	0.009	0.000	25.210	0.046	0.014	0.002	27.221	0.061	0.032	0.062	26.034
Age squared/100	-0.110	0.017	0.000	6.861	-0.072	0.026	0.006	8.008	-0.107	0.059	0.075	7.309
No education	-0.103	0.009	0.000	0.080	-0.085	0.032	0.017	0.294	0.050	0.061	0.399	0.322
Primary	-0.112	0.009	0.000	0.785	-0.020	0.019	0.290	0.606	-0.007	0.035	0.843	0.556
Secondary + (*reference*)												
Urban X No education	0.019	0.028	0.464	0.047	0.017	0.045	0.694	0.228	-0.126	0.024	0.000	0.022
Literacy rate, age 15+ (TA)												
Mean population age (TA)					-0.028	0.016	0.076	20.939				
% in poverty (TA)	0.001	0.0001	0.000	56.613					0.0004	0.0002	0.188	51.142
% in ultra-poverty (TA)					-0.0001	0.0002	0.705	37.580				
% of orphans (age <=14), (TA)	0.005	0.001	0.000	9.120	0.026	0.011	0.020	8.223	0.054	0.004	0.000	10.311
Mean age at first sex (TA)	-0.040	0.002	0.000	15.902	-0.102	0.038	0.008	17.062				
Mean age at first marriage, women (TA)									-0.085	0.011	0.000	18.742
Share of women who report knowledge of HIV prevention methods (TA)					-0.524	0.302	0.082	0.938				
obs. P	0.141				0.170				0.228			
pred. P (at x-bar)	0.126				0.151				0.212			
Log likelihood	-752.674				-1111.669				-1311.518			
Pseudo R2	0.057				0.061				0.060			
N of obs.	1960				2594				2600			

Note: standard errors adjusted for clustering on sentinel site

Table 5: Observed (2001 sentinel), predicted (2001 sentinel) and predicted (2000 DHS) HIV prevalence among women age 15-49, by region/district, %

	Observed (2001 sentinel survey)		Predicted (2001 sentinel survey)	*Predicted, SAE methodology (2000 DHS)*		
Region/district	N	Mean	Mean	N	Mean	Std. Err.
North						
Chitipa				190	8.9	*1.5*
Karonga	188	12.2	14.1	941	9.6	*1.1*
Mzimba	759	16.6	17.4	781	8.6	*1.4*
Nkhata Bay	485	18.4	19.1	186	9.5	*1.1*
Rumphi	528	13.8	14.4	89	14.2	*1.5*
Total (region)	*1,960*	*15.9*	*16.7*	*2,187*	***9.8***	*1.3*
Central						
Dedza	155	3.9	5.5	497	5.4	*1.6*
Dowa	149	4.7	4.8	447	2.1	*0.5*
Kasungu	153	5.2	8.8	728	8.5	*1.7*
Lilongwe	596	20.0	21.6	871	11.1	*1.3*
Mchinji	506	23.5	25.1	340	8.5	*1.4*
Nkhotakota	525	18.9	22.6	221	10.4	*1.3*
Ntcheu	510	18.6	18.9	435	8.9	*1.9*
Ntchisi				185	4.1	*1.0*
Salima				784	7.2	*0.9*
Total (region)	*2,594*	*17.5*	*19.3*	*4,508*	***8.2***	*1.3*
South						
Balaka				226	11.7	*1.7*
Blantyre	623	28.6	30.5	1,023	20.4	*1.4*
Chikwawa				337	12.9	*1.2*
Chiradzulu	201	15.9	26.3	253	22.1	*1.9*
Machinga	131	13.0	13.7	798	15.7	*2.7*
Mangochi	526	16.0	23.2	654	17.9	*2.9*
Mulanje	459	24.6	25.5	905	25.6	*2.5*
Mwanza				121	11.0	*1.5*
Nsanje	475	35.8	39.4	188	19.3	*1.9*
Phalombe				239	22.9	*2.9*
Thyolo	185	17.3	18.4	882	16.5	*2.4*
Zomba				899	13.4	*1.6*
Total (region)	*2,600*	*24.1*	27.7	*6,525*	*18.0*	*1.1*
Total (country)	7,154	19.4	*21.4*	*13,220*	**13.2**	*1.0*

Note: all descriptive statistics involving DHS uses DHS sample weights

DHS-based predictions for each group of women are based on estimating the model which uses group-specifc weights making the distribution by age/gender in the sentinel survey reflect that in the DHS

Table 6: Observed (2004 DHS), and predicted (2000 DHS) HIV prevalence among women age 15-49, selected districts

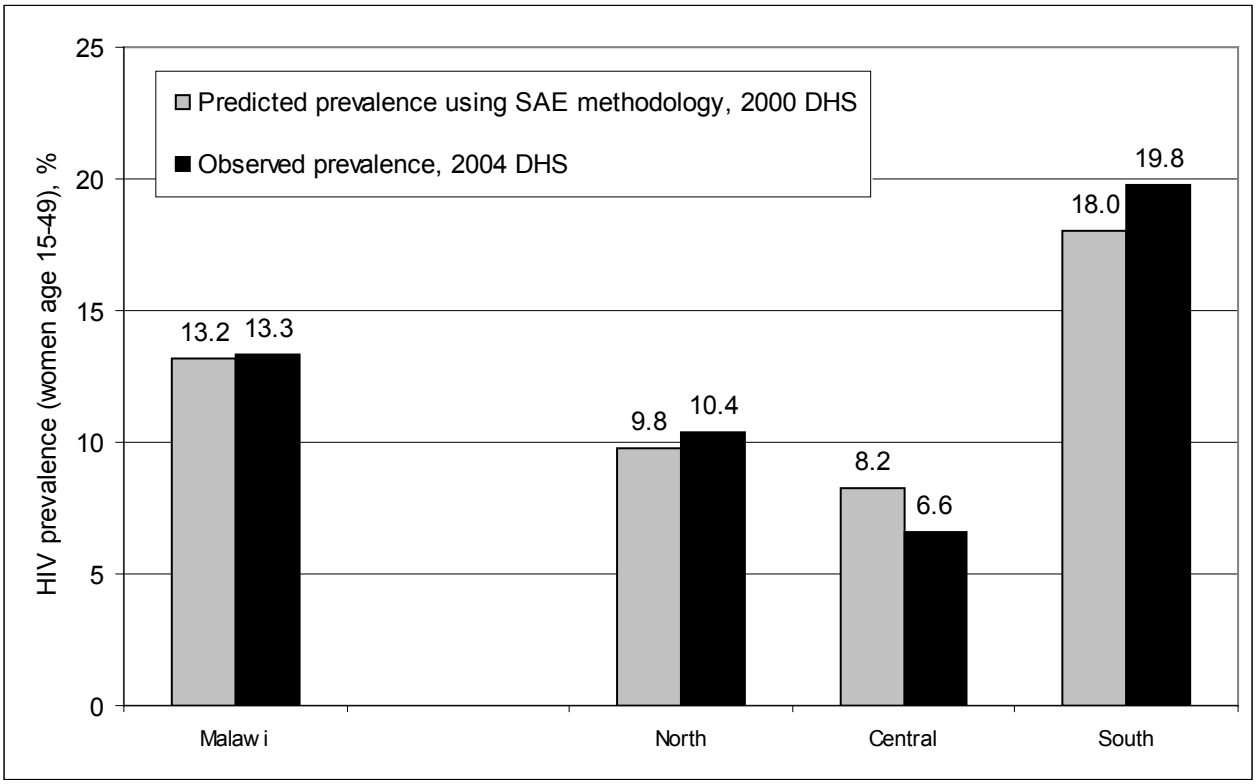
District	*Observed (2004 DHS)*	*N (2004 DHS)*	*Predicted, SAE methodology (2000 DHS)*	*Std Err (partial)*
Blantyre	22.5	211	20.4	1.4
Kasungu	5.5	116	8.5	1.7
Machinga	14.9	99	15.7	2.7
Mangochi	21.4	136	17.9	2.9
Mzimba	6.4	178	8.6	1.4
Salima	9.5	74	7.2	0.9
Thyolo	23.1	145	16.5	2.4
Zomba	24.6	134	13.4	1.6
Mulanje	23.3	117	25.6	2.5
Total (country)	13.3	2,686	13.2	1.0

Table 7: Predicted HIV prevalence using SAE methodology vs. observed and adjusted HIV prevalence using the 2004 Malawi DHS, women age 15-49

	Observed, 2001 sentinel survey	***Predicted, using SAE methodology, 2000 DHS***			Observed, 2004 DHS			Adjusted (for non-response), 2004 DHS		
Geographic Area	Prevalence	Prevalence	95% confidence interval		Prevalence	95% confidence interval		Prevalence	95% confidence interval	
Malawi, excluding Lilongwe	17.8	13.5	11.8	15.2	15.1	13.8	16.4	14.8	13.8	15.8
Lilongwe	20.0	11.1	5.5	16.7	1.6	0.0	4.2	11.5	10.0	13.1
Malawi total	19.4	13.2	11.5	14.9	13.3	12.1	14.6	14.4	13.5	15.3

Figures

Figure 1: Predicted HIV prevalence using SAE methodology vs. observed prevalence in the 2004 DHS, national and regional (women age 15-49)



Appendix

Table A1. The list of commune/district level variables which are likely to be correlated with the probability of HIV infection

Source of data/variables	Unit of measurement	Existing level of aggregation	Level of aggregation used in the analysis	Rationale for the variable
Health facilities data				
Share of health facilities (HF) with antenatal care (ANC) services	%	district	district	A proxy for availability of information to women about HIV/AIDS
Share of HF with family planning services	%	district	district	A proxy for availability of information about the ways of HIV prevention, safe sexual practices, and availability of condoms
Share of HF with pharmacies	%	district	district	A proxy for availability of condoms
Share of HF with HIV/AIDS testing services	%	district	district	A proxy for the extent of spread of information about HIV (which is likely to influence behavior – people are more aware of their own or someone's else HIV status)
Share of HF with HIV/AIDS counseling services	%	district	district	A larger share probably indicates that a higher proportion of HIV infected will have information on what to do not to infect a partner
Share of HF with STD treatment services	%	district	district	A proxy for STD treatment rate; STD treatment decreases the chances of HIV infection; a proxy for information about HIV transmission modes and safe sexual practices
Census data				
Poverty	% of poor and ultra-poor (less than 60% below the poverty line); depth of poverty	TA (traditional authority/ administrative ward)	TA	Poverty level at the TA/district level is likely to be related to the poverty status of the household and the individual, which can transform into sexual behavior in a myriad of ways (affordability of condoms, number of partners, commercial sex)
Population density	People/sq. km.	TA	TA	This variable captures more isolated areas, urban/rural differences in income levels, cultural values, access to sex

				services
Difference in annual population growth rates between 1977-1998, and 1987-1998	Percentage points	TA	TA	A proxy for the intensity of cross-border migration along Mozambique border (as a result of 1987 civil war in Mozambique); a more negative value indicates a larger flow of immigrants from Mozambique (*note: this is likely to be a better variable than a simple dummy for districts bordering Mozambique, and other countries*)
Mean age of population	years	TA	TA	Younger population generally resides in the urban centers and in the northern and central regions (the effect of this variable on the risk of HIV a priori is not clear; you need to see if there is any substantial variation in this variable); it may capture variations in fertility levels (higher fertility as a result of more unprotected sex)
Mean household size		TA	TA	Reflects the differences in fertility and in household stability (smaller households are found in areas where the population follows matrilineal rules of kinship)
Mean number of children born per woman (for women age 12+)		TA	TA	A proxy for fertility rate (central regions have higher fertility rates); probably can serve as a proxy for the use of condoms (?), although there are many other methods of preventing pregnancy (but not infection)
Religion	Christian, Islam, other, none (%)	EA (census enumeration area)/TA	TA	The distribution by religion in the district is likely to affect people's sexual behavior through religious and social values
Type of toilet	Flush, latrine, pit, none (%)	EA (census enumeration area)	TA	A proxy for the level of wellbeing (poverty) in the district, which is likely to be correlated with individual wellbeing (and thus behavior)
Tenure (ownership of the dwelling)	Owner/family occupied, rented, other (%)	EA (census enumeration area)	TA	A large share of rented dwellings can be an indicator of high proportion of SAEsonal labor migrants, refugees, etc. – an environment with a higher prevalence of HIV (and thus of the higher risk of infection)
Share of orphans (defined for children age 14 and less who have at least one parent dead(	%	EA (census enumeration area), TA	TA	HIV prevalence rates affect the number of orphans through the death rates of parents
Nationality (the country where the person come from)	Malawi, Zambia, Tanzania,	EA (census enumeration area)	TA	Districts (communities) which are subject to cross-border moves are more likely to have HIV prevalence different from other districts (depending on the HIV prevalence in other

	Mozambique, Zimbabwe, India, South Africa (%)			countries; refugees from Mozambique are more likely to experience the loss of families, family ties, social values, and engage in risky sexual behaviors (or be exploited) – this will increase the risk of HIV infection for "native" Malawi people
Age structure of the population (5-year intervals)	N. of people in 5-year intervals	EA (census enumeration area)	TA	A share of prime-age men and women is likely to be positively related to the HIV prevalence rate
Mean age at first marriage, women	years	TA	TA	A lower marriage age is likely to decreases the chances of infection (due to the limited number of partners), but a lot depends on the sexual behavior of the partner
Age at first marriage, gender differences	years	TA	TA	Larger gaps in rural areas may indicate barriers to early marriage for men through higher bride-wealth requirements (increased chance of infection as a partner has more partners before getting married)
Literacy rate, age 15+	%	TA	TA	A proxy for the diffusion of knowledge about ways of decreasing the risk of HIV infection, and for the likelihood of people in the region adjusting their sexual behavior so that the risk of infection could be reduced
Economically active population	%	TA	TA	This variable is related to many other factors (income levels, poverty, affordability of condoms, etc.) and can have either a positive or negative effect on the risk of HIV infection
Mean distance to roads (either all, or primary and secondary only)	kilometers	TA	TA	A proxy of mobility (and extent of interactions with people outside of their local area) – a higher mobility is generally associated with the higher risk of HIV infection
GIS data				
Distance to the nearest health facility	meters	EA	EA/TA	A proxy for the availability of information about methods of lowering the risk of HIV infection, availability of condoms
Elevation	meters above SAE level	EA	EA/TA	A proxy for how remote the population point is from "civilized" world
Additive amount of lights per geographic area	density units	EA	EA/TA	May be a good indicator of population density/infrastructure, although most of the values are zero
DHS data				
Age at first sex	years	individual	TA (TA mean is constructed from cluster means; cluster	Early age at first sex increases chances of unprotected sex, and increases the length of exposure to sex (incl. unprotected sex)

			corresponds to EA)	
Knowledge of condom	%	individual	TA	Knowledge of condom
Knowledge of a place to get a method of FP	%	individual	TA	A proxy for knowledge and availability of FP in the area
Number of sexual partners in the last 12 months		individual	TA	Higher number of partners increases the risk of HIV infection in the case of unprotected sex
Share of women who knowledge a place to get condom	%	individual	TA	A proxy for availability of condoms in TA
Share of women who could get a condom if necessary	%	individual	TA	Another proxy for availability of condoms in TA
Share of women who believe something can be done to avoid the risk of HIV infection	%	individual	TA	A proxy for the share of women with "correct" believes (which is likely to transform into less risky behavior)
Share of women who believe condom use can prevent from HIV infection	%	individual	TA	A proxy for the share of women with "correct" believes (which is likely to transform into less risky behavior)
Share of women who believe that it is possible for a healthy-looking person to be HIV infected	%	individual	TA	A proxy for the share of women with "correct" believes (which is likely to transform into less risky behavior)
Share of women who have ever been tested for HIV/AIDS	%	individual	TA	A larger share of people who knows their HIV status is likely to get reflected into HIV prevention behaviors

Note: TA (traditional authority/ administrative ward) refers to the lowest level of administrative division in Malawi. TA means obtained from the Census data will be more accurate than TA means based on DHS cluster means since the former are based on the larger number of observations. District means obtained from TA means should thus also be more accurate than district means obtained from DHS cluster means for the same reason.